A Method for Service Identification of SSL/TLS Encrypted Traffic with the Relation of Session ID and Server IP

Sung-Min Kim, Young-Hoon Goo and Myung-Sup Kim
Dept. of Computer and Information Science
Korea University
Korea
{gogumiking, gyh0808, tmskim}@korea.ac.kr

Soo-Gil Choi, Mi-Jung Choi
Dept. of Computer Science
Kangwon National University
Korea
sgchoi0529@hotmail.com, mjchoi@kangwon.ac.kr

***Abstract*—The SSL/TLS, one of the most popular encryption protocol, was developed as a solution of various network security problem while the network traffic has become complex and diverse. But the SSL/TLS traffic has been identified as its protocol name, not its used services, which is required for the effective network traffic management. This paper proposes a new method to generate service signatures automatically from SSL/TLS payload data and to classify network traffic in accordance with their application services. We utilize the certificate publication information field in the certificate exchanging record of SSL/TLS traffic for the service signatures, which occurs when SSL/TLS performs Handshaking before encrypt transmission. We proved the performance and feasibility of the proposed method by experimental result that classify about 95% SSL/TLS traffic with about 90% accuracy for every SSL/TLS services.**

Keywords—SSL/TLS; SSL/TLS Handshake; Traffic Identification; Session ID; Server IP address;

I. INTRODUCTION

Network traffic classification is an essential step for providing stable network services and efficient network resource management, because of becoming complex of network traffic by diffusion of high-speed Internet and development of network equipment. And security problems, those are personal information leakage, invasion of privacy, and steal account, grew all the more serious. To solve these security problem, various encryption protocols were developed like SSL/TLS [1] and SSH [2]. The encrypted traffic is increasing, but the encrypted traffic has been identified mainly as just its protocol name, not its used services. Which is require for the effective network traffic management.

In this paper, we propose a method to identify SSL/TLS encrypted network traffic in accordance with their application services with firstly with payload signature-based network traffic identification method [3]. In addition, we propose a method that identify unidentified traffic with the relation of session ID and server IP.

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(2015R1D1A3A01018057) and by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT & Future Planning (2010-0020728).

This paper describes in the following order. Second section explains SSL/TLS protocol and SSL/TLS handshake protocol, and third section describes related works to identifying network traffic. Fourth section describes the payload signature/session ID-Server IP based method for service identification of SSL/TLS encrypted traffic. Fifth section describes result of experiment with identification system described in third session. And final section describes conclusion and future works.

II. SSL/TLS PROTOCOL

This section describes SSL/TLS protocol and SSL/TLS handshake that uses for SSL/TLS service identification. The SSL/TLS protocol, an encryption technology, was developed by Netscape for secure data transmission between web servers in 1994. The SSL/TLS traffic is unit of records, like TCP is unit of packets. SSL/TLS record protocol is a hierarchical structure, and each class include record's type, SSL/TLS version, data length, and real data.

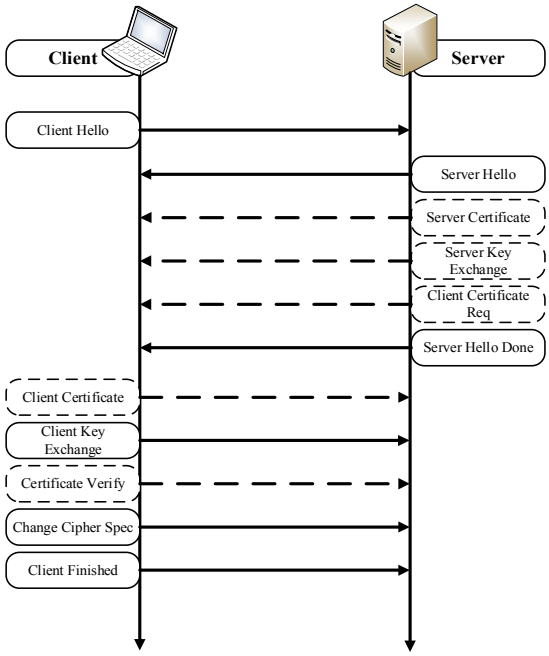
SSL/TLS makes a session with handshake [4] before server and client communicate by encrypted data. By this process, server and client exchange certificate, key, encryption algorithm and so on, and certify each other. Figure 1 illustrates how to SSL/TLS handshake protocol work on network between server and client. First of all client send ClientHello message to server for trying to connect and send its usable cipher suite.

The server, reseived hello message from client, reply ServerHello, ServerCertificate, and ServerHelloDone messages to client. ServerHello message is a response to ClientHello message, and at that time, the session ID is generated and sent. The set of session ID and client address is stored in server's session ID table for a certain period of time. The ServerCertificate message sends certificate with public key. If Client requests to do not send certificate, Server does not send certificate. ServerKeyExchange message is used when it is not enough with certificate exchanging. The CertificateRequest message requests client's certificate, if server needs additory certification. Finally ServerHelloDone

message notify client of ClientHello end and client needs to start request.

If client receives request of client certificate from server, client sends certificate by ClientCertificate message. Since then exchange their key by ClientKeyExchange message and validate client's certificate by CertificateVerify message.

Fig. 1. SSL/TLS full/abbreviated handshake protocol



And SSL/TLS abbreviated handshake protocol works excludes dotted lined things of figure 1. When Client try to reconnect with session ID, gotten from server before, there is no necessity for making session by full handshake. If server remembers the session ID with its session ID table, they can make a session faster without certificate exchanging, key exchanging and so on. On the other hand, if session ID is deleted from server's session ID table, server send alert message to client and they reconnect by full handshake.

III. Related Work

The various network traffic service identifying methods are evolving with development of network technique. Those methods have different merits and demerits, those Header-based, Signature-based and Statistics-based Identifying method are typical cases. This section describes the various traffic identifying methods.

A. *Signature-based Identifying Method*

The signature-based identifying method [5] that matches up signatures with payload directly. This method guarantee high identification rate and accuracy because of matching up signature with payload directly. But it takes high overload and slow processing speed. For overcoming this weakness, some researches was done but it still waste lots of time and effort to generate signature by manual labor. And frequent renewal work required because of short lifecycle of Internet based applications. Also it is hard to generate signatures of encrypted network traffic.

B. *Statistics-based Identifying Method*

The final method is statistics-based identifying method [6] that identifies traffic with statistical information of flow like packet size and packet inter-arrival time. This method can identify fast and encrypted traffic, because it do not check the payload directly. A research, identifying SSL/TLS traffic [7] with this method, was done. The research identified SSL/TLS traffic with 22 numbers of statistical information by various machine-learning based algorithms like "AdaBoost", "C4.5" and "Naïve Bayes". But the research did not get out of limits, dependent on specific network and cannot identify detail as services.

IV. Method for Service Identification of SSL/TLS Encrypted Traffic

This section describes the method for service identification of SSL/TLS encrypted traffic(SSIM : SSL/TLS Identification Method) with human readable payload signature and the relation of session ID and server IP, and define a service as a server, providing network service.

Before identify SSL/TLS application service with payload signature or relation of server IP and session ID, each of services' signatures should be generated. We define the certificate publication information field in the certificate exchanging record of SSL/TLS traffic for the service's signatures, which occurs when SSL/TLS performs handshaking before encrypt transmission.

Fig. 2. Overall system for Service Identification of SSL/TLS Encrypted Traffic with the relation of Server IP and Session ID

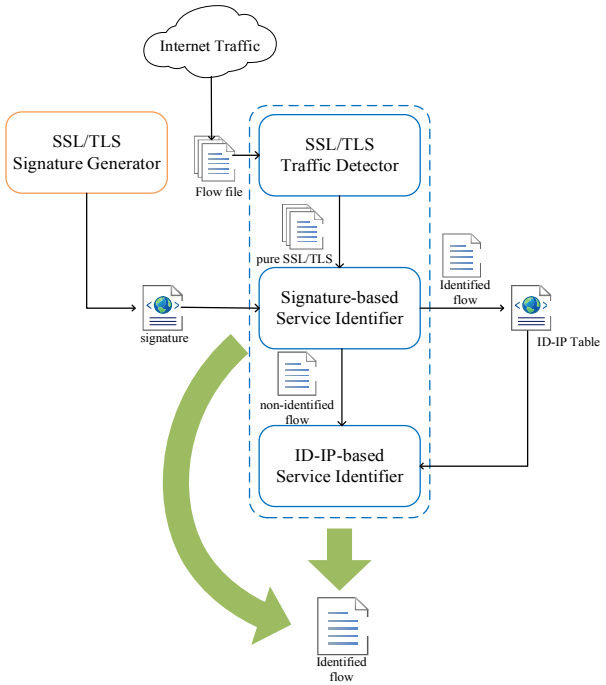


Figure 2 is a process diagram of overall system for service identification of SSL/TLS encrypted traffic with the relation of

server IP and session ID. This system is made up three modules those SSL/TLS Traffic Detector, Signature-based Service Identifier, and ID-IP-based Service Identifier. First, the SSL/TLS Traffic Detector filters pure SSL/TLS traffic before service identification of SSL/TLS traffic. This process is for identification speed improvement by except for non SSL/TLS traffic as an identification target. The Signature-based Service Identifier identifies the Certificate authority information field in the certificate exchanging record as SSL/TLS services with generated signatures before, and record that flow's Server IP and Session ID for next identification. Finally, ID-IP-based Service Identifier does additory identification. It identifies non-identified flow from Signature-based Service Identifier with relation of server IP and session ID.

A. SSL/TLS Traffic Detector

It takes long time for identifying traffic. So before service identifying, some process is necessary. The first of them detects SSL/TLS protocol level and passes only pure SSL/TLS traffic to next service identification step. This process is for fast identifying by excluding non-SSL/TLS traffic from target subject of analysis.

The detecting module is simple that just check front part of payload with SSL/TLS header. This header is defined with 5 bytes, the first 1 byte for type, next 2 bytes for SSL/TLS version and final 2 bytes for data length behind header. SSL/TLS has four types of protocol, their codes are 20, 21, 22, 23 those mean ChangeCipherSpec, Alert, Handshake and ApplicationData each. The latest SSL/TLS version is SSL 3.3(TLS 1.2) and its' code is converted '0303' with SSL version as version field. The length field defines data length behind length field by Big-Endian. All of SSL/TLS packets start this pattern. So if SSL/TLS Traffic Detector works any packets of the flow, includes the packet, can be defined SSL/TLS traffic.

B. Signature-based Service Identifier

The defined traffic as SSL/TLS by SSL/TLS Traffic Detector becomes the input of the Signature-based Service Identifier. It identifies SSL/TLS traffic as a service with payload signature. It is hard to identify the encrypted data with payload signature, but SSL/TLS traffic can be identified as service because handshaking process is not encrypted.

When server sends Certificate message to client, Certificate message includes information about server from the certificate publication. So SSL/TLS Signature Generator can generate signatures from certificate publication information field. Also the Signature-based Service Identifier identifies services by comparison between the signature and that field.

But not all of SSL/TLS traffic can be identified by this module. Because abbreviated handshake, when client send to server ClientHello message with session ID for reconnection, does not have Certificate message. So this step records sets of server IP and Session ID from identified flows, and the sets are used for additory identifying at next step.

C. ID-IP-based Service Identifier

SSL/TLS session is disconnected, if client do not send any response for period of time. And server and client reconnects by abbreviated handshake without Certificate and Key Exchanging messages. So flows, connected by abbreviated handshake, cannot be identified only by Signature-based Service Identifier which matches signatures to publication information of Certificate message. When client request reconnection to server, client send session ID, received previous connection, on ClientHello message. This session ID is made by server. Therefore this session ID and server IP set is surely a unique set. This module identifies unidentified traffic identification by Signature-based Service Identifier using the unique sets.

The input data of ID-IP-based Service Identifier is unidentified flows and session ID and server IP sets of identified flows. The merit of this module is that can identify SSL/TLS service fast by matching only first packet of flow. The reason is if a flow reconnects by abbreviated handshake, client requests reconnection with session ID by ClientHello message and the ClientHello message is the first of handshake.

These three modules work complementary each other. SSL/TLS Traffic Detector lessen next identifiers' load by detecting only SSL/TLS traffic. Signature-based Service Identifier identifies SSL/TLS traffic as specific services and stores its session ID and server IP sets. Finally, ID-IP-based Service Identifier identifies unidentified SSL/TLS traffic from Signature-based Service Identifier with session ID and server IP sets. There are results of some experimentations for proving performance of these three modules next section.

V. EVALUATION

This section verifies the architecture of the SSL/TLS service identification system. We performed two types of experimentations. The first test is how much SSL/TLS traffic appear on general network. The second experimentation identified SSL/TLS service by generated signatures before. In addition, we compared between traffic identification by only Signature-based Service Identifier and with ID-IP-based Service Identifier. For clear experimentation, other than TCP traffic and flows those have no real data with incorrect handshaking were excluded. Therefore the experimentations used only TCP traffic.

A. The Quantity of SSL/TLS traffic

This test is for how much SSL/TLS traffic appear on general network. The dataset for experimentation was generated for one day from campus.

TABLE 1. SSL/TLS DETECTION

	Flows	Packets	Bytes
Total	6,043,560	1,080,739,271	930,080,255,327
SSL/TLS	506,412	59,164,220	46,224,581,249
Rate	8.38%	5.47%	4.97%

Table 1 shows quantity of SSL/TLS traffic on campus for a day. This result is generated by SSL/TLS Traffic Detector. The total number of flows are about $60 \cdot 10^5$, and SSL/TLS traffic of total flows are $5 \cdot 10^5$. In other word, SSL/TLS traffic generated about 8.38% for a day. It is not too much amount but it is an importance to identify and manage increasing SSL/TLS traffic because present network trend has been changing to SSL/TLS.

B. SSL/TLS Service Identification

For this experimentation, we generated some signatures from three applications, Google, Facebook and Kakaotalk. Google is the most famous searching engine in the world. Facebook is a Social network Service on web and mobile. Kakaotalk is the most popular messenger in Korea. It services on PC application and mobile either. Naturally, these three applications are encrypted by SSL/TLS protocol. And we collected each three applications' traffic, using them randomly.

We identified the traffic with only Signature-based Service Identifier and using ID-IP-based Service Identifier additionally.

TABLE 3. ACCURACY OF SERVICE IDENTIFICATION

Service		Flows	Packets	Bytes
Google	Total	736	69,739	54,823,801
	Identified	599	62,881	50,022,915
	Accuracy	81.39%	90.17%	91.24%
Facebook	Total	264	69,854	67,494,176
	Identified	155	60,071	60,520,703
	Accuracy	58.71%	86.00%	89.67%
Kakaotalk	Total	145	6,912	4,832,639
	Identified	104	5,646	4,421,168
	Accuracy	71.72%	81.68%	91.49%

Table 3 is accuracy of service identification which used only Signature-based Service Identifier. It identified 58.71%~81.39% of flows. The accuracy does not satisfy used only Signature-based Service Identifier. Especially, it did not identify large amount of Facebook. Thus we experimented for increasing the accuracy by using ID-IP-based Service Identifier in addition next.

Fig. 3. Accuracy of Service Identification

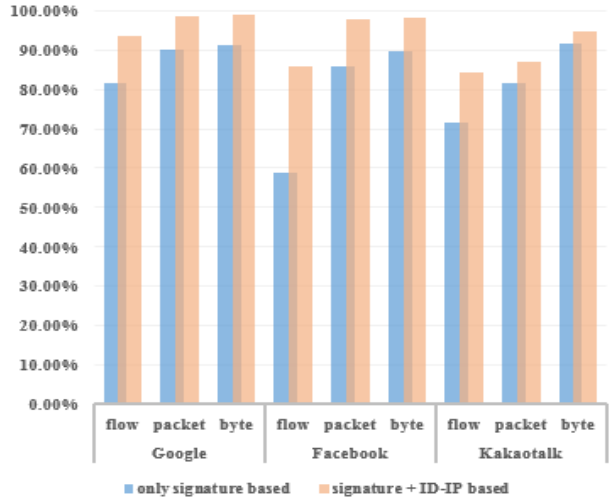


Figure 3 shows increasing of the experimentation. We gained satisfied result from this experimentation that all three services was identified more, if identify with ID-IP-based Service Identifier. It identified 84%~93% of flows, especially identified Facebook up to 27% of flows more. The most of un-identified traffic is alert message of server do not have the session ID on its session ID table. However, if we identified SSL/TLS traffic in real time consistently, we could identify most of these case. If we make policy of expiration time of Session ID and Server IP set in ID-IP table future, the method can work effectively in real time service identification. And the higher accuracy of identification as packet or byte then flow shows the reason of un-identified traffic. Because alert message has less packets and data. We proved SSIM's effectivity and possibility of encryption traffic identification by these experimentations' result.

VI. CONCLUSION AND FUTURE WORK

It is hard to identify encrypted traffic because of its impossibility of expectation. There was various research of classifying SSL/TLS traffic but they could not get out of limits, could not identify as application services.

In this paper, we proposed a method for service identification of SSL/TLS encrypted traffic and developed a system based on the method. The system is built up of SSL/TLS Detector, Signature-based Service Identifier and ID-IP-based Service Identifier. SSL/TLS Detector detects only SSL/TLS traffic and Signature-based Service Identifier identifies SSL/TLS service, matching payload directly with signatures. Finally, ID-IP-based Service Identifier identifies non-identified traffic from Signature-based Service Identifier with set of Session ID and Server IP. And we proved the method by three types of experimentation, it made a result increasing maximum 27% of accuracy.

In future, we plan to make policy of ID-IP table for using storage effectively and increasing identifying speed of the system. It may be removing of expired session ID and server IP sets. And we plan to research other SSL/TLS fields for signatures. By these ways, we are going to improve the performance of SSIM.

REFERENCES

[1] Elgohary, Ashraf, Tarek S. Sobh, and Mohammed Zaki. "Design of an enhancement for SSL/TLS protocols." computers & security 25.4 (2006): 297-306.

[2] Kyoung-Lyoon Kim, Myung-Sup Kim, and Hyoung-joong Kim. "SSH Traffic Identification Using EM Clustering" The Journal of Korea Information and Communications Society 37.12 (2012): 1160-1167.

[3] Jun-Sang Park, Sung-Ho Yoon, Youngjoon Won, and Myung-Sup Kim, "A Lightweight Software Model for Signature-Based Application-Level Traffic Classification System." IEICE TRANSACTIONS on Information and Systems 97.10 (2014): 2697-2705.

[4] Qi, Fang, et al. "Batching SSL/TLS handshake improved." Information and Communications Security. Springer Berlin Heidelberg, 2005. 402-413.

[5] Jun-Sang Park, Sung-Ho Yoon, and Myung-Sup Kim. "Performance Improvement of the Payload Signature based Traffic Classification System Using Application Traffic Locality" The Journal of Korea Information and Communications Society 38.7 (2013): 519-525.

[6] Hyun-Min An, Jae-Hyun Ham, Myung-Sup Kim, "Performance Improvement of the Statistical Information based Traffic Identification System", KIPS Transactions on Computer and Communication Systems(KTCCS) vol.2, No.8, pp.335-342, Aug, 2013

[7] McCarthy Curtis, and A. Nur Zincir-Heywood. "An investigation on identifying SSL traffic." Computational Intelligence for Security and Defense Applications (CISDA), 2011 IEEE Symposium on. IEEE, 2011.